Reinforcement Learning in Rich-Observation MDPs using Spectral Methods

Kamyar Azizzadenesheli
University of California, Irvine
kazizzad@uci.edu

Alessandro Lazaric
SequeL, INRIA Lille, France
alessandro.lazaric@inria.fr

Animashree Anandkumar
University of California, Irvine
a.anandkumar@uci.edu

Abstract

Designing effective exploration-exploitation algorithms in Markov decision processes (MDPs) with large state-action spaces is the main challenge in reinforcement learning (RL). In fact, the learning performance degrades with the number of states and actions in the MDP. However, MDPs often exhibit a low-dimensional latent structure in practice, where a small hidden state is observable through a possibly large number of observations. In this paper, we study the setting of rich-observation Markov decision processes (ROMDP), where hidden states are mapped to observations through an injective mapping, so that an observation can be generated by only one hidden state. While this mapping is unknown a priori, we introduce a spectral decomposition method that consistently estimates how observations are clustered in the hidden states. The estimated clustering is then integrated into an optimistic algorithm for RL (UCRL), which operates on the smaller clustered space. The resulting algorithm proceeds through phases and we show that its per-step regret (i.e., the difference in cumulative reward between the algorithm and the optimal policy) decreases as more observations are clustered together and finally, matches the (ideal) performance of an RL algorithm running directly on the hidden MDP.

1 Introduction

Reinforcement learning (RL) framework studies the agent-environment interaction, in which the agent learns to maximize a given reward function in the long run [6, 23]. At the beginning of the interaction, the agent is uncertain about the environment's dynamics and must *explore* different policies in order to gain information about it. Once the agent is fairly certain, the knowledge about the environment can be *exploited* to compute a good policy attaining a large cumulative reward. Designing algorithms that achieve an effective trade-off between exploration and exploitation is the primary goal of reinforcement learning. The trade-off is commonly measured in terms of *cumulative regret*, that is the difference between the rewards accumulated by the optimal policy (which requires exact knowledge of the environment) and those obtained by the learning algorithm.

In practice, we often deal with environments with large observation spaces (e.g., robotics). In this case the regret tends to be very large as it increases with the size of the observations. Nonetheless, in many domains there is an underlying low dimensional latent space that summarizes the large observation space and its dynamics and rewards (e.g., in robot navigation high-dimensional visual and sensory input can be summarized in 2D as locations). If the mapping between the latent states and the observations is known a priori, then it is trivial to exploit it and design a learning algorithm that operates on the low-dimensional latent space rather than the large observation space. However, this mapping is typically unknown and needs to be learnt by the agent. It is then crucial to develop efficient learning algorithms that can quickly discover the mapping between hidden and observation states and fully exploit the small size of the hidden space.

Contributions. In this paper we focus on rich-observation Markov decision processes (ROMDP), where a small number of X hidden states are mapped to a large number of Y observations through an injective mapping, so that an observation can be generated by only one hidden state and hidden states can be viewed as clusters. If the mapping was known, one could directly run, e.g., UCRL and obtained a regret scaling only with the number of hidden states X instead of the number of observations. More formally, the regret is reduced from $\widetilde{O}(D_Y Y\sqrt{N})$ down to $\widetilde{O}(D_X X\sqrt{N})$, where D is the diameter of the MDP (i.e., the longest shortest path between any two observations/states) and N is the total number of steps over which the regret is cumulated.

In this setting, we show that it is indeed possible to devise an algorithm that starting from observations can progressively cluster them in "smaller" states and eventually converge to the hidden MDP. We introduce SL-UC, where we integrate spectral decomposition methods into the upper-bound for RL algorithm (UCRL) [13]. The algorithm proceeds in epochs in which an estimated mapping between observations and hidden state is computed and an optimistic policy is computed on the MDP (called auxiliary MDP) constructed from the samples collected so far and the estimated mapping. The mapping is computed using spectral decomposition of the tensor associated to the observation process. We prove that this method is guaranteed to correctly "cluster" observations together with high probability. As a result, the dimensionality of the auxiliary MDP decreases as more observations are clustered, thus making the algorithm more efficient computationally and more effective in finding good policies. Under suitable assumptions, we derive a regret bound showing that the per-step regret decreases over epochs and we prove a worst-case bound on the number of steps (and corresponding regret) before the full mapping between states and observations is computed. The regret cumulated over this period is actually constant as the time to correct clustering does not increase with the number of steps N. As a result, we have that SL-UC asymptotically matches the regret of learning directly on the latent MDP. We also notice that the improvement in the regret comes with an equivalent reduction in time and space complexity. In fact, as more observations are clustered, the space to store the auxiliary MDP decreases and the complexity of the extended value iteration step in UCRL decreases from $O(Y^3)$ down to $O(X^3)$.

Related work. The assumption of the existence of a latent space is often used to reduce the learning complexity. For multi-armed bandits, Gheshlaghi-Azar et al. [10] and Maillard and Mannor [19] assume that a bandit problem is generated from an unknown (latent) finite set and show how the regret can be significantly reduced by learning this set. Gentile et al. [9] consider the more general scenario of latent contextual bandits, where the contexts belong to a few underlying hidden classes. They show that a uniform exploration strategy over the contexts, combined with an online clustering algorithm achieve a regret scaling only with the number of hidden clusters. An extension to recommender systems is considered in [11] where the contexts for the users and items are unknown a priori. Again, uniform exploration is used together with the spectral algorithm of Anandkumar et al. [1] to learn the latent classes.

The ROMDP model considered is a generalization of the latent contextual bandits, where actions influence the contexts (i.e., the states) and the objective is to maximize the long-term reward. ROMDPs have been studied in [16] in the PAC-MDP setting and episodic deterministic environments using an algorithm searching the best Q-function in a given function space. This result is extended to the general class of contextual decision processes in [14]. Ortner [21] proposes an algorithm integrating state aggregation with UCRL but, while the resulting algorithm may significantly reduce the computational complexity of UCRL, the analysis does not show any improvement in the regret. Finally, we notice that ROMDPs are a special class of partially observable MDPs (POMDP). Azizzadenesheli et al. [4] recently proposed an algorithm that leverages spectral methods to learn the hidden dynamic of POMDPs and derived a regret scaling as $\sqrt{Y}$ using fully stochastic policies (which are sub-optimal in ROMDPs). While the computation of the optimal memoryless policy relies on an optimization oracle, which in general is NP-hard [3, 18], computing the optimal policy in ROMDPs amounts to solving a standard MDP. Finally, Guo et al. [12] develops a PAC-MDP analysis for learning in episodic POMDPs and obtain a bound that depends on the size of the observations.

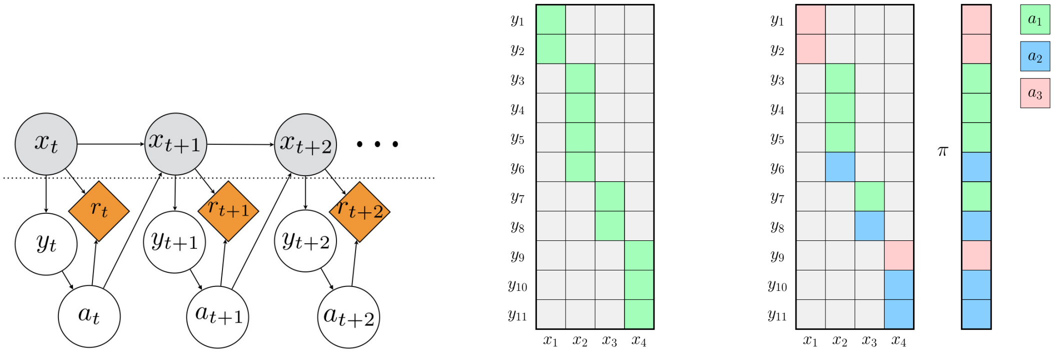

Figure 1: *(left)* Graphical model of a ROMDP. *(middle)* Example of an observation matrix O. Since state and observation labelling is arbitrary, we arranged the non-zero values so as to display a diagonal structure. *(right)* Example of clustering that can be achieved by policy π (e.g., $\mathcal{X}_\pi^{(a_1)} = \{x_2, x_3\}$). Using each action we can recover *partial* clusterings corresponding to 7 auxiliary states $\mathcal{S} = \{s_1..s_7\}$ with clusters $\mathcal{Y}_{s_1} = \{y_1, y_2\}$ $\mathcal{Y}_{s_2} = \{y_3, y_4, y_5\}$, and $\mathcal{Y}_{s_8} = \{y_{10}, y_{11}\}$, while the remaining elements are the singletons y_6, y_7, y_8, and y_9.

2 Rich Observation MDPs

A rich-observation MDP (ROMDP) (Fig. 1(*left*)) is a tuple $M = \langle \mathcal{X}, \mathcal{Y}, \mathcal{A}, R, f_T, f_O \rangle$, where $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{A}$ are the sets of hidden states, observations, and actions. We denote by X, Y, A their cardinality and we enumerate their elements by $i \in [X] = \{1..X\}$, $j \in [Y] = \{1..Y\}$, $l \in [A] = \{1..A\}$. We assume that the hidden states are much fewer than the observations, i.e., $X \ll Y$. We consider rewards bounded in $[0, 1]$ that depend only on hidden states and actions with a reward matrix $R \in \mathbb{R}^{A \times X}$ such that $[R]_{i,l} = \mathbb{E}[r(x = i, a = l)]$. The dynamics of the MDP is defined on the hidden states as $T_{i',i,l} := f_T(i'|i, l) = \mathbb{P}(x' = i'|x = i, a = l)$, where $T \in \mathbb{R}^{X \times X \times A}$ is the transition tensor. The observations are generated as $[O]_{j,i} = f_O(j|i) = \mathbb{P}(y = j|x = i)$, where the observation matrix $O \in \mathbb{R}^{Y \times X}$ has minimum *non-zero* entry $O_{\min}$. This model is a strict subset of POMDPs since each observation y can be generated by only one hidden state (see Fig. 1-*middle*) and thus $\mathcal{X}$ can be seen as a non-overlapping clustering of the observations.

We denote by $\mathcal{Y}_x = \mathcal{Y}_i = \{y = j \in \mathcal{Y} : [O]_{j,i} > 0\}$ the set of observations in cluster x, while $x_y = x_j$ is the cluster observation $y = j$ belongs to.[1] This structure implies the existence of an *observable* MDP $M_{\mathcal{Y}} = \langle \mathcal{Y}, \mathcal{A}, R', f'_T \rangle$, where $R' = R$ as the reward of an observation-action pair (y, a) is the same as in the hidden state-action pair (x_y, a), and the dynamics can be obtained as $f'_T(j'|j, a) = \mathbb{P}(y' = j'|y = j, a = l) = \mathbb{P}(y' = j'|x' = x_{j'})\mathbb{P}(x' = x_{j'}|x = x_j, a = l) = [O]_{j',x_{j'}}[T]_{x_{j'},x_j,l}$. We measure the performance of an observation-based policy $\pi_{\mathcal{Y}} : \mathcal{Y} \to \mathcal{A}$ starting from a hidden state x by its asymptotic average reward $\rho(x; \pi_{\mathcal{Y}}) = \lim_{N\to\infty} \mathbb{E}\big[\sum_{t=1}^N r_t/N \big| x_0 = x, \pi_{\mathcal{Y}}\big]$. Given the mapping between the ROMDP to the hidden MDP, the optimal policy $\pi^*_{\mathcal{Y}}(y)$ is equal to the optimal hidden-state policy $\pi^*_{\mathcal{X}} : \mathcal{X} \to \mathcal{A}$ for all $y \in \mathcal{Y}_x$. The learning performance of an algorithm run over N steps is measured by the regret

$$R_N = N\rho^* - \Big[\sum_{t=1}^{N} r_t\Big], \textit{ where } \rho^* = \rho(\pi^*_{\mathcal{X}}).$$

Finally we recall that the diameter of the observation MDP is defined as $D_{\mathcal{Y}} = \max_{y,y' \in \mathcal{Y}} \min_{\pi:\mathcal{Y}\to\mathcal{A}} \mathbb{E}\big[\tau_\pi(y, y')\big]$, where $\tau_\pi(y, y')$ is the (random) number of steps from y to y' by following the observation-based policy π (similar for the diameter of the hidden MDP).

3 ROMDP Recovery Through Spectral Methods

In this section we introduce the spectral method used to to learn the structure of the observation matrix O. In particular, we show that we do not need to estimate O exactly as the clusters $\{\mathcal{Y}_x\}_{x\in\mathcal{X}}$ can be recovered by identifying the non-zero entries of O. We need a first assumption on the ROMDP.

Assumption 1. *The Markov chain induced on the hidden MDP M by any policy $\pi_{\mathcal{Y}}$ is ergodic.*

Under this assumption for any policy π there is a stationary distribution over hidden states ω_π and a stationary distribution conditional on an action $\omega_\pi^{(l)}(i) = \mathbb{P}_\pi(x = i|a = l)$. Let $\mathcal{X}_\pi^{(l)} = \{i \in [X] :$

[1] Throughout the paper we use the indices i, j, and l and the "symbolic" values x, y, and a interchangeably.

$\omega_\pi^{(l)}(i) > 0\}$ be the hidden states where action l could be taken according to policy π. In other words, if $\mathcal{Y}_\pi^{(l)} = \{j \in [Y] : \pi(j) = l\}$ is the set of observations in which policy π takes action l, then $\mathcal{X}_\pi^{(l)}$ is the set of hidden states $\{x_y\}$ with $y \in \mathcal{Y}_\pi^{(l)}$ (see Fig. 1-*right*). We also define the set of all hidden states that can be reached starting from states in $\mathcal{X}_\pi^{(l)}$ and taking action l, that is

$$\overline{\mathcal{X}}_\pi^{(l)} = \bigcup_{i \in \mathcal{X}_\pi^{(l)}} \Big\{ i' \in [X] : \mathbb{P}\big(x' = i' | x = i, a = l\big) > 0 \Big\}.$$

Similarly $\underline{\mathcal{X}}_\pi^{(l)}$ is the set of hidden states from which we can achieve the states $\mathcal{X}_\pi^{(l)}$ by policy π. We need the following assumption.

Assumption 2 (Full-Rank)**.** *Given any action l, the slice of transition tensor $[T]_{\cdot,\cdot,l}$ is full rank.*

Asm. 2 implies that for any action l the dynamics of M is "expansive", i.e., $|\mathcal{X}_\pi^{(l)}| \leq |\overline{\mathcal{X}}_\pi^{(l)}|$. In other words, the number of hidden states where policy π can take an action l (i.e., $\mathcal{X}_\pi^{(l)}$) is smaller than the number of states that can be reached when executing action l itself (i.e., $\overline{\mathcal{X}}_\pi^{(l)}$).

Multi-view model and exact recovery. We are now ready to introduce the multi-view model [1] that allows us to reconstruct the clustering structure of the ROMDP. We consider the trajectory of observations and actions generated by an arbitrary policy π and we focus on three consecutive observations y_{t-1}, y_t, y_{t+1} at any step t. As customary in multi-view models, we *vectorize* the observations into three one-hot view vectors $\vec{v}_1$, $\vec{v}_2$, $\vec{v}_3$ in $\{0,1\}^Y$ such that $\vec{v}_1 = \vec{e}_j$ means that the observation in the first view is $j \in [Y]$ and where we remap time indices $t-1, t, t+1$ onto 1, 2, and 3. We notice that these views are indeed independent random variables when conditioning on the state x_2 (i.e., the hidden state at time t) and the action a_2 (i.e., the action at time t), thus defining a multi-view model for the hidden state process. Let $k_1 = |\underline{\mathcal{X}}_\pi^{(l)}|$, $k_2 = |\mathcal{X}_\pi^{(l)}|$ and $k_3 = |\overline{\mathcal{X}}_\pi^{(l)}|$, then we define the factor matrices $V_1^{(l)} \in \mathbb{R}^{Y \times k_1}$, $V_2^{(l)} \in \mathbb{R}^{Y \times k_2}$, $V_3^{(l)} \in \mathbb{R}^{Y \times k_3}$ as follows

$$[V_p^{(l)}]_{j,i} = \mathbb{P}(\vec{v}_p = \vec{e}_j | x_2 = i, a_2 = l), \text{ with } i \in \underline{\mathcal{X}}_\pi^{(l)} \text{ for } p{=}1, i \in \mathcal{X}_\pi^{(l)} \text{ for } p{=}2, i \in \overline{\mathcal{X}}_\pi^{(l)} \text{ for } p{=}3.$$

We are interested in estimating $V_2^{(l)}$ since it directly relates to the observation matrix as

$$[V_2^{(l)}]_{j,i} = \frac{\mathbb{P}(a_2 = l | y_2 = j)\mathbb{P}(y_2 = j | x_2 = i)}{\mathbb{P}(a_2 = l | x_2 = i)} = \frac{\mathbb{I}\{\pi(j) = l\} f_O(j|i)}{\mathbb{P}(a_2 = l | x_2 = i)}, \tag{1}$$

where $\mathbb{I}$ is the indicator function. As it can be noticed, $V_2^{(l)}$ borrows the same structure as the observation matrix O and since we want to recover only the clustering structure of M (i.e., $\{\mathcal{Y}_i\}_{i\in[X]}$), it is sufficient to compute the columns of $V_2^{(l)}$ up to any multiplicative constant. In fact, any non-zero entry of $V_2^{(l)}$ corresponds to a non-zero element in the original observation matrix (i.e., $[V_2^{(l)}]_{j,i} > 0 \Rightarrow [O]_{j,i} > 0$) and for any hidden state i, we can construct a cluster $\mathcal{Y}_i^{(l)} = \{j \in [Y] : [V_2^{(l)}]_{j,i} > 0\}$, which is accurate up to a re-labelling of the states. More formally, there exists a mapping function $\sigma^{(l)} : \mathcal{X} \to \mathcal{X}$ such that any pair of observations $j, j' \in \mathcal{Y}_i^{(l)}$ is such that $j, j' \in \mathcal{Y}_{\sigma(i)}$. Nonetheless, as illustrated in Fig. 1-*right*, the clustering may not be minimal. In fact, we have $[O]_{j,i} > 0 \not\Rightarrow [V_2^{(l)}]_{j,i} > 0$ since $[V_2^{(l)}]_{j,i}$ may be zero because of policy π, even if $[O]_{j,i} > 0$. Since the (unknown) mapping function $\sigma^{(l)}$ changes with actions, we are unable to correctly "align" the clusters and we may obtain more clusters than hidden states. We define $\mathcal{S}$ as the auxiliary state space obtained by the partial aggregation and we prove the following result.

Lemma 1. *Given a policy π, for any action l and any hidden state $i \in \mathcal{X}_\pi^{(l)}$, let $\mathcal{Y}_i^{(l)}$ be the observations that can be clustered together according to $V_2^{(l)}$ and $\mathcal{Y}^{\mathsf{c}} = \mathcal{Y} \setminus \bigcup_{i,l} \mathcal{Y}_i^{(l)}$ be the observations not clustered, then the auxiliary state space $\mathcal{S}$ contains all the clusters $\{\bigcup_{i,l} \mathcal{Y}_i^{(l)}\}$ and the singletons in $\mathcal{Y}^{\mathsf{c}}$ for a total number of elements $S = |\mathcal{S}| \leq AX$.*

We now show how to recover the factor matrix $V_2^{(l)}$. We introduce mixed second and third order moments as $K_{p,q}^{(l)} = \mathbb{E}[\vec{v}_p \otimes \vec{v}_q]$, $K_{p,q,r}^{(l)} = \mathbb{E}[\vec{v}_p \otimes \vec{v}_q \otimes \vec{v}_r]$ where p, q, r is any permutation of $\{1,2,3\}$. Exploiting the conditional independence of the views, the second moments can be written as $K_{p,q}^{(l)} = \sum_{i \in \mathcal{X}_\pi^l} \omega_\pi^{(l)}(i)[V_p^{(l)}]_{:,i} \otimes [V_q^{(l)}]_{:,i}$ where $[V_p^{(l)}]_{:,i}$ is the i-th column of $V_p^{(l)}$. In general the

Algorithm 1 Spectral learning algorithm.

```
Input: Trajectory $(y_1, a_1, \ldots, y_N)$
For Action $l \in [A]$ do
  Estimate second moments $\widehat{K}^{(l)}_{2,3}$, $\widehat{K}^{(l)}_{1,3}$, $\widehat{K}^{(l)}_{2,1}$, and $\widehat{K}^{(l)}_{3,1}$
  Estimate the rank of matrix $\widehat{K}^{(l)}_{2,3}$ (see App. B)
  Compute symmetrized views $\widetilde{v}_{1,t}$ and $\widetilde{v}_{3,t}$, for $t = 2..N-2$
  Compute second and third moments $\widehat{M}^{(l)}_2$ and $\widehat{M}^{(l)}_3$
  Compute $\widehat{V}^{(l)}_2$ from the tensor decomposition of (an orthogonalized version of) $\widehat{M}^{(l)}_3$
  return clusters
      $\widehat{\mathcal{Y}}^{(l)}_i = \{j \in [Y] : [\widetilde{V}^{(l)}_2]_{j,i} > 0\}$
```

Algorithm 2 Spectral-Learning UCRL(SL-UC).

```
Initialize: $t = 1$, initial state $x_1$, $k = 1$, $\delta/N^6$
While $t < N$ do
  Run Alg. 1 on samples from epoch $k-1$ and obtain $\widehat{\mathcal{S}}$
  Compute aux. space $\widehat{\mathcal{S}}^{(k)}$ by merging $\widehat{\mathcal{S}}$ and $\widehat{\mathcal{S}}^{(k-1)}$
  Compute the estimate reward $r^{(k)}$ and dynamics $p^{(k)}$
  Construct admissible AuxMDPs $\mathcal{M}^{(k)}$
  Compute the optimistic policy
      $\widetilde{\pi}^{(k)} = \arg\max_{\pi} \max_{M \in \mathcal{M}^{(k)}} \rho(\pi; M)$        (6)
  Set $v^{(k)}(s,l) = 0$ for all actions $l \in \mathcal{A}$, $s \in \widehat{\mathcal{S}}^{(k)}$
  While $\forall l, \forall s, v^{(k)}(s,l) < \max\{1, N^{(k)}(s,l)\}$ do
    Execute $a_t = \widetilde{\pi}^{(k)}(s_t)$
    Observe reward $r_t$ and observation $y_t$
```

second moment matrices are rank deficient, with rank $X^{(l)}_\pi$. We can construct a symmetric second moment by introducing the symmetrized views

$$\widetilde{v}_1 = K^{(l)}_{2,3}(K^{(l)}_{1,3})^{\dagger}\vec{v}_1, \qquad \widetilde{v}_3 = K^{(l)}_{2,1}(K^{(l)}_{3,1})^{\dagger}\vec{v}_3, \tag{2}$$

where $K^{\dagger}$ denotes the pseudoinverse. Then we can construct the second and third moments as

$$M^{(l)}_2 = \mathbb{E}\big[\widetilde{v}_1 \otimes \widetilde{v}_3\big] = \sum_{i \in \mathcal{X}^{(l)}_\pi} \omega^{(l)}_\pi(i)[V^{(l)}_2]_{:,i} \otimes [V^{(l)}_2]_{:,i}. \tag{3}$$

$$M^{(l)}_3 = \mathbb{E}\big[\widetilde{v}_1 \otimes \widetilde{v}_3 \otimes \vec{v}_2\big] = \sum_{i \in \mathcal{X}^{l}_\pi} \omega^{(l)}_\pi(i)[V^{(l)}_2]_{:,i} \otimes [V^{(l)}_2]_{:,i} \otimes [V^{(l)}_2]_{:,i}. \tag{4}$$

We can now employ the standard machinery of tensor decomposition methods to orthogonalize the tensor $M^{(l)}_3$ using $M^{(l)}_2$ and recover $V^{(l)}_2$ (refer to [1] for further details) and a suitable clustering.

Lemma 2. *For any action $l \in [A]$, let $M^{(l)}_3$ be the third moment constructed on the symmetrized views as in Eq. 4, then we can orthogonalize it using the second moment $M^{(l)}_2$ and obtain a unique spectral decomposition from which we compute the exact factor matrix $[V^{(l)}_2]_{j,i}$. As a result, for any hidden state $i \in \mathcal{X}^{(l)}_\pi$ we define the cluster $\widetilde{\mathcal{Y}}^{(l)}_i$ as*

$$\widetilde{\mathcal{Y}}^{(l)}_i = \{j \in [Y] : [V^{(l)}_2]_{j,i} > 0\} \tag{5}$$

and there exists a mapping $\sigma^{(l)} : X \to X$ such that if $j, j' \in \widetilde{\mathcal{Y}}^{(l)}_i$ then $j, j' \in \mathcal{Y}_{\sigma^{(l)}(i)}$ (i.e., observations that are clustered together in $\widetilde{\mathcal{Y}}^{(l)}_i$ are clustered in the original ROMDP).

Spectral learning. While in practice we do not have the exact moments, we can only estimates them through samples. Let N be the length of the trajectory generated by policy π, then we can construct $N-2$ triples $\{y_{t-1}, y_t, y_{t+1}\}$ that can be used to construct the corresponding views $\vec{v}_{1,t}, \vec{v}_{2,t}, \vec{v}_{3,t}$ and to estimate second mixed moments as $\widehat{K}^{(l)}_{p,q} = \frac{1}{N(l)}\sum_{t=1}^{N(l)-1}\mathbb{I}(a_t = l)\ \vec{v}_{p,t} \otimes \vec{v}_{q,t}$, with $p, q \in \{1,2,3\}$ and $N(l) = \sum_{t}^{N-1}\mathbb{I}(a_t = l)$. Furthermore, we require knowing $|\mathcal{X}^{(l)}_\pi|$, which is not known apriori. Under Asm. 1 and 2, for any action l, the rank of $K^{(l)}_{2,3}$ is indeed $|\mathcal{X}^{(l)}_\pi|$ and thus $\widehat{K}^{(l)}_{2,3}$ can be used to recover the rank. The actual way to calculate the efficient rank of $\widehat{K}^{(l)}_{2,3}$ is quite intricate and we postpone the details to App. B. From $\widehat{K}^{(l)}_{p,q}$ we can construct the symmetric views $\widetilde{v}_{1,t}$ and $\widetilde{v}_{3,t}$ as in Eq. 2 and compute the estimate second and third

Lemma 3. *Under Asm. 1 and 2, let $\widehat{V}^{(l)}_2$ be the empirical estimate of $V^{(l)}_2$ obtained using N samples generated by a policy π. There exists N_0 such that for any $N(l) > N_0$, $l \in \mathcal{A}$, $i \in \mathcal{X}^{(l)}_\pi$ w.p. $1-\delta$*

$$\|[V^{(l)}_2]_{\cdot,i} - [\widehat{V}^{(l)}_2]_{\cdot,i}\|_2 \le C_2\sqrt{\frac{\log(2Y^{3/2}/\delta)}{N(l)}} := \mathcal{B}^{(l)}_O \tag{7}$$

where C_2 is a problem-dependent constant independent from the number of observations Y.

While this estimate could be directly used to construct a clustering of observations, the noise in the empirical estimates might lead to $[\widehat{V}_2^{(l)}]_{j,i} > 0$ for any (j,i) pair, which prevents us from generating any meaningful clustering. On the other hand, we can use the guarantee in Lem. 3 to single-out the entries of $\widehat{V}_2^{(l)}$ that are non-zero w.h.p. We define the binary matrix $\widetilde{V}_2^{(l)} \in \{0,1\}^{Y \times X}$ as

$$[\widetilde{V}_2^{(l)}]_{j,i} = \begin{cases} 1 & \text{if } [\widehat{V}_2^{(l)}]_{j,i} \geq \mathcal{B}_O^{(l)} \\ 0 & \text{otherwise} \end{cases},$$

which relies on the fact that $[\widehat{V}_2^{(l)}]_{j,i} - \mathcal{B}_O^{(l)} > 0$ implies $[V_2^{(l)}]_{j,i} > 0$. At this point, for any l and any $i \in \mathcal{X}_\pi^{(l)}$, we can generate the cluster $\widehat{\mathcal{Y}}_i^{(l)} = \{j \in [Y] : [\widetilde{V}_2^{(l)}]_{j,i} > 0\}$, which is guaranteed to aggregate observations correctly in high-probability. We denote be $\widehat{\mathcal{Y}}^{\mathrm{c}} = \mathcal{Y} \setminus \bigcup_{i,l} \widehat{\mathcal{Y}}_i^{(l)}$ the set of observations which are not clustered through this process. Then we define the auxiliary state space $\widehat{\mathcal{S}}$ obtained by enumerating all the elements of non-clustered observations together with clusters $\{\widehat{\mathcal{Y}}_i^{(l)}\}_{i,l}$, for which we have the following guarantee.

Corollary 1. *Let $\widehat{\mathcal{S}}$ be the auxiliary states composed of clusters $\{\widehat{\mathcal{Y}}_i^{(l)}\}$ and singletons in $\mathcal{Y}^{\mathrm{c}}$ obtained by clustering observations according to $\widetilde{V}_2^{(l)}$, then for any pair of observations j, j' clustered together in $\widehat{\mathcal{S}}$, there exists a hidden state i such that $j, j' \in \mathcal{Y}_i$. Finally, $\widehat{\mathcal{S}} \rightarrow \mathcal{S}$ as N tends to infinity.*

4 Spectral Learning UCRL

We now describe the spectral learning UCRL (SL-UC) (Alg. 2) obtained by integrating the spectral method above with the UCRL strategy. The learning process is split into epochs of increasing length. At the beginning of epoch k, we use the trajectory $(s_1, a_1, .., s_{N^{(k-1)}})$ generated at previous epoch using auxiliary states $s \in \widehat{\mathcal{S}}^{(k)}$ to construct the auxiliary state space $\widehat{\mathcal{S}}$ using Alg. 1[2]. As discussed in the previous section, the limited number of samples and the specific policy executed at epoch $k-1$ may prevent from clustering many observations together, which means that despite $\widehat{\mathcal{S}}$ being a *correct* clustering (see Cor. 1), its size may still be large. While clusterings obtained at different epochs cannot be "aligned" because of different labelling, we can still effectively merge together any two clusterings $\widehat{\mathcal{S}}$ and $\widehat{\mathcal{S}}'$ generated by two different policies π and π'. We illustrate this procedure through Fig. 2(*left*). Observations y_3, y_4, and y_5 are clustered together in the auxiliary space generated by π, while y_5 and y_6 are clustered together using π'. While the labelling of the auxiliary states is arbitrary, observations preserve their labels across epochs and thus we can safely conclude that observations y_3, y_4, y_5, and y_6 belong to the same hidden state. Similarly, we can construct a new cluster with y_9, y_{10}, and y_{11}, which, in this case, returns the exact hidden space $\mathcal{X}$. Following this procedure we generate $\widehat{\mathcal{S}}^{(k)}$ as the clustering obtain by merging $\widehat{\mathcal{S}}$ and $\widehat{\mathcal{S}}^{(k-1)}$ (where $\widehat{\mathcal{S}}^1 = \mathcal{Y}$).

At this point we can directly estimate the reward and transition model of the auxiliary MDP constructed on $\widehat{\mathcal{S}}^{(k)}$ by using standard empirical estimators. Let $N^{(k)}(y,a,y')$ be the number of times a transition from observation y to y' through action a has been observed up to the beginning of epoch k. Then we can easily construct $N^{(k)}(s,a,s') = \sum_{y \in s} \sum_{y' \in s'} N^{(k)}(y,a,y')$, where with an abuse of notation we write $y \in s$ to denote the fact that observation y has been clustered into an auxiliary state s at epoch k. Similarly we can compute the number of visits to any auxiliary state-action pair $N^{(k)}(s,a)$ and the reward cumulated over time $R^{(k)}(s,a)$. Then we return the estimates[3] $\widehat{r}^{(k)}(s,a) = R^{(k)}(s,a)/N^{(k)}(s,a)$, $\widehat{p}^{(k)}(s'|s,a) = N^{(k)}(s,a,s')/N^{(k)}(s,a)$. The corresponding confidence intervals are such that for any $s \in \widehat{\mathcal{S}}^{(k)}$ and $a \in \mathcal{A}$

[2] Since Alg. 1 receives as input a sequence of auxiliary states rather than observations as in Sect. 3 the spectral decomposition runs on a space of size $|\widehat{\mathcal{S}}^{(k-1)}|$ instead of Y, thus reducing the computation complexity.

[3] Since the clustering $\widehat{\mathcal{S}}^{(k)}$ is *monotonic* (i.e., observations clustered at epoch k stay clustered at any other epoch $k' > k$), $\widehat{r}^{(k)}$ and $\widehat{p}^{(k)}$ can be computed incrementally without storing the statistics $N^{(k)}(y,a,y')$, $N^{(k)}(y,a)$, and $R^{(k)}(y,a)$ at observation level, thus significantly reducing the space complexity of the algorithm.

$$\|p(\cdot|s,a) – \widehat{p}^{(k)}(\cdot|s,a)\|_1 \leq d_p(s,a) = \sqrt{\frac{7S^{(k)}\log(\frac{2AN^{(k)}}{\delta})}{N^{(k)}(s,a)}},$$

$$|\bar{r}(s,a) - \widehat{r}^{(k)}(s,a)| \leq d_r(s,a) = \sqrt{\frac{7\log(\frac{2S^{(k)}AN^{(k)}}{\delta})}{2N^{(k)}(s,a)}},$$

hold w.p. $1-\delta$, where $p(\cdot|s,a)$ and $\bar{r}$ are the transition probabilities and reward of the auxiliary MDP $M_{\widehat{\mathcal{S}}^{(k)}}$. At this point we can simply apply the same steps as in standard UCRL, where an optimistic auxiliary MDP $\widetilde{M}^{(k)}$ is constructed using the confidence intervals above and extended value iteration (EVI) [13]. The resulting optimal optimistic policy $\widetilde{\pi}^{(k)}$ is then executed until the number samples at least for one pair of auxiliary state and action is doubled. EVI has a per-iteration complexity which scales as $\mathcal{O}((\widehat{S}^{(k)})^2 A)$ thus gradually reducing the complexity of UCRL on the observation space (i.e., $\mathcal{O}((Y)^2 A)$) as soon as observations are clustered together.

Theorem 1. *Consider a ROMDP $M = \langle \mathcal{X}, \mathcal{Y}, \mathcal{A}, R, f_T, f_O \rangle$ with diameter $D_{\mathcal{X}}$. If* SL-UC *is run over N time steps, under Asm. 1 and 2, with probability $1-\delta$ it suffers the total regret of*

$$\textit{Reg}_N \leq \sum_{k=1}^{K} \Big(D_{\widehat{S}^{(k)}} \sqrt{\widehat{S}^{(k)} \log\Big(\frac{N^{(k)}}{\delta}\Big)} \sum_{s \in \widehat{\mathcal{S}}^{(k)}, a} \frac{\nu^{(k)}(s,a)}{\sqrt{N^{(k)}(s,a)}} \Big), \tag{8}$$

where $(\mathcal{S}^{(k)})$ is the sequence of auxiliary state spaces generated over K epochs.

Remark. This bound shows that the per-step regret decreases over epochs. First we notice that only the regret over the first few (and short) epochs actually depends on the number of observations Y and the diameter $D_{\mathcal{Y}}$. As soon as a few observations start being clustered into auxiliary states, the regret depends on the number of auxiliary states $\widehat{S}^{(k)}$ and the diameter $D_{\mathcal{S}^{(k)}}$. Since $\widehat{S}^{(k)}$ decreases every time an observation is added to a cluster and $D_{\mathcal{S}^{(k)}}$ is monotonically decreasing with of $\widehat{S}^{(k)}$, the per-step regret significantly decreases with epochs.[4] Cor. 1 indeed guarantees that the number of auxiliary states in $\widehat{\mathcal{S}}$ reduces down to $|\mathcal{S}|$ (XA in the worst case) as epochs get longer. Furthermore we recall that even if the clustering $\widehat{\mathcal{S}}$ returned by the spectral method is not minimal, merging clusters across epochs may rapidly result in very compact representations even after a few epochs.

Minimal clustering. While Thm. 1 shows that the performance of SL-UC improves over epochs, it does not relate it to the (ideal) performance that could be achieved when the hidden space had been known. In order to provid a minimal clustrting, we integrate Alg. 2 with a clustering technique similar to the one used in [9] and [21]. At any epoch k, we proceed by merging together all the auxiliary states in $\widehat{\mathcal{S}}^{(k)}$ whose reward and transition confidence intervals overlap (i.e., s and s' are merged if the confidence interval $[\widehat{r}(s,a) \pm d_r(s,a)]$ overlaps with $[\widehat{r}(s',a) \pm d_r(s',a)]$ and $[\widehat{p}(\cdot|s,a) \pm d_p(s,a)]$[5] overlaps with $[\widehat{p}(\cdot|s',a) \pm d_p(s',a)]$. If the number of new clusters is equal to X, then we claim we learned the true clustering, if it is less than X we ignore this temporary clustering and proceed to the next epoch. It worth to note that this procedure requires the knowledge of X, while the spectral method, by its own, does not. While an explicit rate of clustering is very difficult to determine (the merging process depends on the spectral method, whose result depends on the policy, which in turn is determined according to the clustering at previous epochs), we derive worst-case bounds on the number of steps needed to start clustering at least one observation (i.e., steps before avoiding the dependency on Y and $D_{\mathcal{Y}}$) and before the exact clustering is recovered.

Corollary 2. *Let $\tau_M = \max_{x,\pi} \mathbb{E}_\pi[\tau_\pi(x,x)]$ the maximum expected returning time in MDP M (bounded due to ergodicity) and*

$$\overline{N}_{\textit{first}} = \frac{AY\tau_M}{O_{\min}} \frac{C_2 \log(1/\delta)}{\max_{i,j} f_O(y=j|x=i)^2}; \;\; \overline{N}_{\textit{last}} = \frac{AY\tau_M}{O_{\min}^3} C_2 \log(1/\delta). \tag{9}$$

After $\overline{N}_{\textit{first}}$ steps at least two observations are clustered and after $\overline{N}_{\textit{last}}$ steps all observations are clustered (but not necessarily in the minimum hidden space configuration) with probability $1-\delta$. This implies that after $\overline{N}_{\textit{last}}$ steps $|\widehat{S}^{(k)}| \leq XA$. Furthermore, let $\gamma_r = \min_{x,x',a} |r(x,a) - r(x',a)|$

[4] We refer to the per-step regret since an epoch may be longer, thus making the cumulative epoch regret larger.

[5] Deviation $d_p(s,a)$ on a $\widehat{S}$ dimensional simplex.

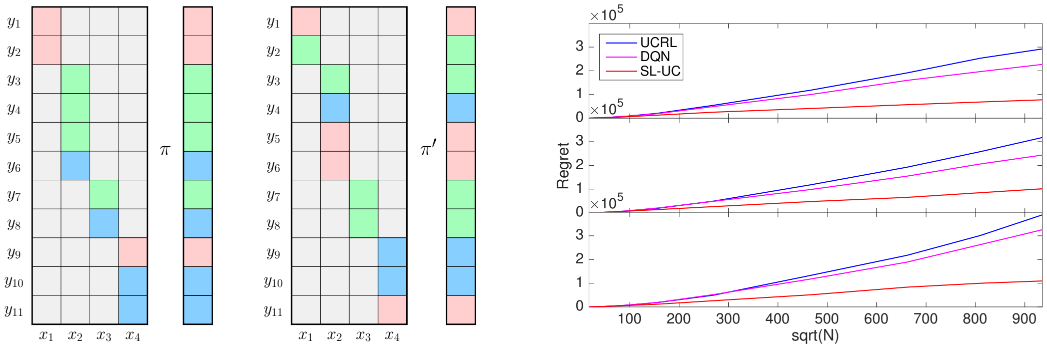

Figure 2: *(left)* Examples of clusterings obtained from two policies that can be effectively merged. *(right)* Regret comparison for ROMDPs with $X = 5, A = 4$ and from top to bottom $Y = 10, 20, 30$.

and $\gamma_p = \min_{x,x',a} \|p(\cdot|x,a) - p(\cdot|x',a)\|_1$ *be the smallest gaps between rewards and transition probabilities and let* $\gamma = \max\{\gamma_r, \gamma_p\}$ *the maximum between the two.*

In worst case, using the additional clustering step together with SL-UC *guarantees that after*

$$\overline{N}_{\mathcal{X}} = \min\Big\{\frac{AY^2\tau_M}{\gamma^2}\log(1/\delta), \max\Big\{\frac{AS^2\tau_M}{\gamma^2}\log(1/\delta), \overline{N}_{last}\Big\}\Big\}$$

samples the hidden state $\mathcal{X}$ *is correctly reconstructed (i.e.,* $\widehat{\mathcal{S}}^{(k)} = \mathcal{X}$*), therefore*

$$Reg_N \le 34 D_{\mathcal{X}} X \sqrt{A(N - \overline{N}_{\mathcal{X}})\log(N/\delta)}\,\mathbb{I}(N \ge \overline{N}_{\mathcal{X}}) + \min\{\overline{N}_{\mathcal{X}}, 34 D_{\mathcal{Y}} Y \sqrt{A(\overline{N}_{\mathcal{X}})\log(N/\delta)}\}$$

We first notice that this analysis is constructed over a series of worst-case steps (see proof in App. E.1). Nonetheless, it first shows that the number of observations Y does impact the regret only over the first $\overline{N}_{\text{first}}$ steps, after which $\widehat{S}^{(k)}$ is already smaller than $\mathcal{Y}$. Furthermore, after at most $\overline{N}_{\text{last}}$ the auxiliary space has size at most XA (while the diameter may still be as large as $D_{\mathcal{Y}}$). Finally, after $\overline{N}_{\mathcal{X}}$ steps $\widehat{S}^{(k)}$ reduces to $\mathcal{X}$ and the performance of SL-UC tends to the same performance of UCRL in the hidden MDP.

5 Experiments

We validate our theoretical results by comparing the performance of SL-UC, UCRL2 (model based, exact solution) and DQN (model free, function approximation) [20], two well known RL algorithms. Since SL-UC is proposed specifically for infinite horizon environment, we are not able to compare it with episodic setting, e.g., PSRL [22]. For DQN we implemented a three hidden-layers feed forward network (with no CNN block), equipped with RMSprop and replay buffer. We tune the hyper parameters of the network and report the best performance achieved by network of size $30 \times 30 \times 30$.

We consider three randomly generated ROMDPs (Dirichlet transition and Uniform reward with different bias) with $X = 5$, $A = 4$ and observation spaces of sizes $Y = 10, 20, 30$. Fig. 2(*right*) reports the regret on a $\sqrt{N}$ scale. We see that in each instance the regret of UCRL and DQN grows much faster than SL-UC's. While all regrets tend to be linear (i.e., growing as $\sqrt{N}$), we clearly see that the performance of UCRL and DQN is negatively affected by the increasing number of observations, while the regret of SL-UC stays almost constant, confirming that the hidden space $\mathcal{X}$ is learned very rapidly. This experimental results are the first step towards more practical applications. We report additional experiments in App. G.

6 Conclusion

We introduced SL-UC, a novel RL algorithm to learn in ROMDPs combining a spectral method for recovering the clustering structure of the problem and UCRL to effectively trade off exploration and exploitation. We proved theoretical guarantees showing that SL-UC progressively refines the clustering so that its regret tends to the regret that could be achieved when the hidden structure is known in advance (in higher order term). We showed that how the SL-UC gradually constructs smaller MDPs, therefore, has lower cost in computing optimistic policy, encounters fewer number of epochs, and suffers from lower computation cost. Finally, this work opens several interesting directions to extend the results for variety of state aggregation topologies [17]. Furthermore, one can aggregate the proposed method with other regret analyses, e.g. [8] and leverage the current bounds.

7 Acknowledgement

K. Azizzadenesheli is supported in part by NSF Career Award CCF-1254106. A. Lazaric is supported in part by a grant from CPER Nord-Pas de Calais/FEDER DATA Advanced data science and technologies 2015-2020, CRIStAL (Centre de Recherche en Informatique et Automatique de Lille), and the French National Research Agency (ANR) under project ExTra-Learn n.ANR-14-CE24-0010-01. A. Anandkumar is supported in part by Microsoft Faculty Fellowship, Google faculty award, Adobe grant, NSF Career Award CCF- 1254106, and AFOSR YIP FA9550-15-1-0221.

References

[1] Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832.

[2] Anandkumar, A., Hsu, D., and Kakade, S. M. (2012). A method of moments for mixture models and hidden markov models. *arXiv preprint arXiv:1203.0683*.

[3] Azizzadenesheli, K., Lazaric, A., and Anandkumar, A. (2016a). Open problem: Approximate planning of pomdps in the class of memoryless policies. *arXiv preprint arXiv:1608.04996*.

[4] Azizzadenesheli, K., Lazaric, A., and Anandkumar, A. (2016b). Reinforcement learning of pomdps using spectral methods. In *Proceedings of the 29th Annual Conference on Learning Theory (COLT)*.

[5] Azizzadenesheli, K., Lazaric, A., and Anandkumar, A. (2017). Experimental results: Reinforcement learning of pomdps using spectral methods. *arXiv preprint arXiv:1705.02553*.

[6] Bertsekas, D. and Tsitsiklis, J. (1996). *Neuro-Dynamic Programming*. Athena Scientific.

[7] Cesa-Bianchi, N., Gentile, C., and Zappella, G. (2013). A gang of bandits. In *Advances in Neural Information Processing Systems*, pages 737–745.

[8] Dann, C., Lattimore, T., and Brunskill, E. (2017). Ubev-a more practical algorithm for episodic rl with near-optimal pac and regret guarantees. *arXiv preprint arXiv:1703.07710*.

[9] Gentile, C., Li, S., and Zappella, G. (2014). Online clustering of bandits. In *ICML*, pages 757–765.

[10] Gheshlaghi-Azar, M., Lazaric, A., and Brunskill, E. (2013). Sequential transfer in multi-armed bandit with finite set of models. In *Advances in Neural Information Processing Systems 26*, pages 2220–2228.

[11] Gopalan, A., Maillard, O.-A., and Zaki, M. (2016). Low-rank bandits with latent mixtures. *arXiv preprint arXiv:1609.01508*.

[12] Guo, Z. D., Doroudi, S., and Brunskill, E. (2016). A pac rl algorithm for episodic pomdps. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 510–518.

[13] Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600.

[14] Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2016). Contextual decision processes with low bellman rank are pac-learnable. *arXiv preprint arXiv:1610.09512*.

[15] Kontorovich, A., Weiss, R., et al. (2014). Uniform chernoff and dvoretzky-kiefer-wolfowitz-type inequalities for markov chains and related processes. *Journal of Applied Probability*, 51(4):1100–1113.

[16] Krishnamurthy, A., Agarwal, A., and Langford, J. (2016). Pac reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, pages 1840–1848.

[17] Li, L., Walsh, T. J., and Littman, M. L. (2006). Towards a unified theory of state abstraction for mdps. In *Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics (ISAIM-06)*.

[18] Littman, M. L. (1994). Memoryless policies: Theoretical limitations and practical results. In *Proceedings of the Third International Conference on Simulation of Adaptive Behavior : From Animals to Animats 3: From Animals to Animats 3*, SAB94, pages 238–245, Cambridge, MA, USA. MIT Press.

[19] Maillard, O.-A. and Mannor, S. (2014). Latent bandits. In *Proceedings of the Thirty-First International Conference on Machine Learning (ICML'14)*.

[20] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

[21] Ortner, R. (2013). Adaptive aggregation for reinforcement learning in average reward markov decision processes. *Annals of Operations Research*, 208(1):321–336.

[22] Osband, I., Russo, D., and Van Roy, B. (2013). (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011.

[23] Sutton, R. S. and Barto, A. G. (1998). *Introduction to reinforcement learning*. MIT Press.

A Proof of Lemma 3

At the end of each epoch, e.g. k, we estimate the factor matrix $V_2^{(l)}$ (for all $l \in [A]$) using all the samples collected during that epoch according to policy $\widetilde{\pi}^{(k)}$. In order to simplify the notation, in the following we remove the dependency on k, even if all the quantities should be intended as specifically computed at the beginning of epoch k.

In order to bound the empirical error of the moment estimators, we need to consider the properties of the Markov chain generated by policy $\widetilde{\pi}$ and the fact that a single continuous trajectory is observed. In particular, we have to carefully consider the mixing time of the underlying Markov chain (the amount of time it takes that the underlying Markov chain converges to its stationary distribution) and exploit the martingale property of the trajectory. This problem has been previously studied by Azizzadenesheli et al. [4] for the general case of partially observable MDPs. Since ROMDPs are a special case of POMDPs, we can directly rely on the following general concentration inequality.

For any ergodic Markov chain with stationary distribution $\omega_{\widetilde{\pi}}$, let $f_{1\rightarrow t}(x_t|x_1)$ by the distribution over states reached by a policy $\widetilde{\pi}$ after t steps starting from an initial state x_1. The inverse mixing time $\rho_{\text{mix},\pi}(t)$ of the chain is defined as

$$\rho_{\text{mix},\widetilde{\pi}}(t) = \sup_{x_1} \|f_{1\rightarrow t}(\cdot|x_1) - \omega_{\widetilde{\pi}}\|_{\text{TV}} ,$$

where $\|\cdot\|_{\text{TV}}$ is the total-variation metric. Kontorovich et al. [15] show that for any ergodic Markov chain the mixing time can be bounded as

$$\rho_{\text{mix},\widetilde{\pi}}(t) \leq G(\widetilde{\pi})\theta^{t-1}(\widetilde{\pi}),$$

where $1 \leq G(\widetilde{\pi}) < \infty$ is the *geometric ergodicity* and $0 \leq \theta(\widetilde{\pi}) < 1$ is the *contraction coefficient* of the Markov chain generated by policy $\widetilde{\pi}$.

Proposition 1 (Theorem 11 [4] POMDP concentration bound)**.** *Consider a sequence of ν observations $\{y_1, \ldots, y_\nu\}$ obtained by executing a policy $\widetilde{\pi}$ in a POMDP starting from an arbitrary initial hidden state. For any action $l \in [A]$, $\nu^{(l)}$-length sequence $b^{(l)} = \{(y_{t-1}, y_t, y_{t+1}); a_t = l\}$, and any c-Lipschitz*[6] *matrix valued function $\Phi(\cdot) : b^{(l)} \rightarrow \mathbb{R}^{Y\times Y}$, we have*

$$\left\|\Phi(b^{(l)}) - \mathbb{E}[\Phi(b^{(l)})]\right\|_2 \leq \frac{G(\widetilde{\pi})}{1-\theta(\widetilde{\pi})}\Big(1 + \frac{1}{\sqrt{2}c(\nu^{(l)})^{\frac{3}{2}}}\Big)\sqrt{8c^2\nu^{(l)}\log\Big(\frac{2Y}{\delta}\Big)}$$

with probability at least $1-\delta$, where $G(\widetilde{\pi})$ and $\theta(\widetilde{\pi})$ are, respectively, the geometric ergodicity and the contraction coefficient of the underlying Markov chain on the hidden states (they define how fast the underlying Markov chain converges to its stationary distribution), and the expectation is with respect to the distribution of initial state equals to the stationary distribution.

The parameters $1 \leq G(\pi^{(k)}) < \infty$ and $0 \leq \theta(\pi^{(k)}) < 1$ are well defined for Markov chain and shows the state distribution of the Markov chain convergence to its stationary distribution (if such distribution exists) with rate of $G(\pi)\theta(\pi)^t$. (the lower $G(\pi)$ and $\theta(\pi)$ give the lower mixing for corresponding Markov chain.) In this paper, we are interested in moments of our data, therefore, $\Phi(\cdot)$ is considered as a moment estimator.

Given the ergodicity assumption (Asm. 1) under any policy, we can apply Proposition 1 to bound the errors for both second and third order moments. For any $\{p, q, r\}$ a permutation of set $\{1, 2, 3\}$

$$\|\widehat{K}^{(l)}_{p,q} - K^{(l)}_{p,q}\|_2 \leq G(\pi)\frac{1 + \frac{1}{\sqrt{2}c(\nu^{(k)}(l))^{\frac{3}{2}}}}{1-\theta}\sqrt{8c^2\nu^{(k)}(l)log(\frac{2Y}{\delta})} \tag{10}$$

$$\|\widehat{M}^{(l)}_{p,q,r} - M^{(l)}_{p,q,r}\|_2 \leq G(\pi)\frac{1 + \frac{1}{\sqrt{2}c(\nu^{(k)}(l))^{\frac{3}{2}}}}{1-\theta}\sqrt{8c^2\nu^{(k)}(l)log(\frac{2Y^{1.5}}{\delta})} \tag{11}$$

with probability at least $1-\delta$. At this point we can proceed with applying the robust tensor power method proposed in [2] to recover $V_2^{(l)}$ and obtain the guarantees of Lemma 5 of Azizzadenesheli

[6] under the Hamming metric

et al. [4] through Proposition 4, where $c = \frac{1}{\nu^{(k)}(l)}$. We report a more detailed version of the statement of Lemma 3.

Lemma 4 (Concentration Bounds)**.** *The robust power method of Anandkumar et al. [2] applied to tensor $\widehat{M}_3^{(l)}$ returns the $X^{(l)}$ columns of matrix $V_2^{(l)}$ with the following confidence bounds*

$$\left\| [V_2^{(l)}]_{(\cdot|i)} - [\widehat{V}_2^{(l)}]_{(\cdot|i)} \right\|_2 \leq \epsilon_3^{(l)} = C_O^l \sqrt{\frac{\log(2Y^{3/2}/\delta)}{\nu^{(l)}}} := \mathcal{B}_O^{(l)} \tag{12}$$

if

$$\nu^{(l)} \geq \overline{N} := \max_{\pi} \left(\frac{4}{\omega_{\widetilde{\pi}_{\min}}^{(l)} \min\limits_{m\in\{1,2,3\}} \{\sigma_{\min}^2(V_m^{(l)})\}} \right)^2 \log(2\frac{(Y^{1.5})}{\delta})\Theta^{(l)} \tag{13}$$

$$\Theta^{(l)} := \max \left\{ \frac{16(X^{(l)})^{\frac{1}{3}}}{C_1^{\frac{2}{3}}(\omega_{\widetilde{\pi}_{\min}}^{(l)})^{\frac{1}{3}}}, 4, \frac{2\sqrt{2}X^{(l)}}{C_1^2\omega_{\min}^{(l)} \min\limits_{m\in\{1,2,3\}} \{\sigma_{\min}^2(V_m^{(l)})\}} \right\}, \tag{14}$$

with probability at least $1-\delta$, where C_1 is a problem-independent constants and $\omega_{\widetilde{\pi}_{\min}}^{(l)} := \min_{i\in\mathcal{X}^{(l)}} \mathbb{P}_{\widetilde{\pi}}(x = i|a = l)$ where the minimization is over non-zero probabilities. In addition, the $\sigma_{\min}(\cdot)$ operator returns the smallest non-zero singular value of its input matrix. The values of the error $\mathcal{B}_O^{(l)}$ under policy $\widetilde{\pi}$ is defined (see Eq.29 of Azizzadenesheli et al. [4])

$$\mathcal{B}_O^{(l)} := G(\widetilde{\pi}) \frac{4\sqrt{2}+2}{(\omega_{\widetilde{\pi}_{\min}}^{(l)})^{\frac{1}{2}}(1-\theta(\widetilde{\pi}))} \sqrt{\frac{\log(2\frac{(2Y)}{\delta})}{\nu^{(l)}}} + \frac{8\widetilde{\epsilon}^{(l)}}{\omega_{\widetilde{\pi}_{\min}}^{(l)}}, \tag{15}$$

where

$$\widetilde{\epsilon}^{(l)} \leq \frac{2\sqrt{2}G(\widetilde{\pi})\frac{2\sqrt{2}+1}{1-\theta(\widetilde{\pi})}\sqrt{\frac{\log(\frac{2(Y^{\frac{3}{2}})}{\delta})}{\nu^{(l)}}}}{((\omega_{\widetilde{\pi}_{\min}}^{(l)})^{\frac{1}{2}} \min\limits_{m\in\{1,2,3\}} \{\sigma_{\min}(V_m^{(l)})\})^3} + \frac{\left(64G(\widetilde{\pi})\frac{2\sqrt{2}+1}{1-\theta(\widetilde{\pi})}\right)}{\min\limits_{m\in\{1,2,3\}} \{\sigma_{\min}^2(V_m^{(l)})\}(\omega_{\widetilde{\pi}_{\min}}^{(l)})^{1.5}} \sqrt{\frac{\log\left(2\frac{Y^{\frac{3}{2}}}{\delta}\right)}{\nu^{(l)}}},$$

We notice that the columns of $V_2^{(l)}$ are all orthogonal (but not orthonormal) since the clusters are non-overlapping and an observation j that can be obtained from a state i cannot be generated by any other state i' (i.e., for any $i \neq i'$, $[V_2^{(l)}]_{:,i}^{\mathsf{T}}[V_2^{(l)}]_{:,i'} = 0$). As a result, Eq. 3 can be seen as an eigendecomposition of $M_2^{(l)}$, where the columns $[V_2^{(l)}]_{:,i}$ are the eigenvectors and $\omega_\pi^{(l)}(i)$ are the eigenvalues. More formally, let $M_2^{(l)} = U\Sigma U^{\mathsf{T}}$ be the eigendecomposition of $M_2^{(l)}$, if all eigenvalues are distinct, the eigenvectors in U can be used to recover $V_2^{(l)}$ up to a mapping function and multiplicative factors. Nonetheless, in general $V_2^{(l)}$ may have eigenvalues with multiplicity and the eigendecomposition of $M_2^{(l)}$ may return a wrong clustering since observations generated by distinct states (and thus with different rewards and dynamics) may be aggregated together. In this case, we have to move to the third order statistics to disambiguate between observations and cluster them properly.

B Rank recovery

Lemma 4 holds when the rank of matrix $V_2^{(l)}$ is known in advance. While this is not the case in practice, here we show how one can estimate the rank $r = \left|\mathcal{X}_{\pi^{(k)}}^{(l)}\right|$ of $V_2^{(l)}$. Given the expansiveness of latent MDP (Asm. 2), we have that for any policy π and any action l, $\left|\mathcal{X}_\pi^{(l)}\right| \leq \left|\overline{\mathcal{X}}_\pi^{(l)}\right|$. The rank of the second moment matrix $K_{2,3}^{(l)}$ is then $\min\{\left|\mathcal{X}_{\pi^{(k)}}^{(l)}\right|, \left|\overline{\mathcal{X}}_{\pi^{(k)}}^{(l)}\right|\} = r$, which also corresponds to the number of non-zero columns in matrix $V_2^{(l)}$. We can then try to estimate r through the estimate second moment $\widehat{K}_{2,3}^{(l)}$, which according to Eq. 10, estimates $K_{2,3}^{(l)}$ up to an additive error $\epsilon_{2,3}^{(l)}$ that decreases as $O(\sqrt{\frac{1}{\nu^{(l)}}})$. This means that the highest perturbation over its singular values is also at most $O(\sqrt{\frac{1}{\nu^{(l)}}})$. We introduce a threshold function $g^{\epsilon}(\nu^{(l)})$ that satisfies the condition

$$\epsilon_{2,3}^{(l)} \leq g^{\epsilon}(\nu^{(l)}) \leq 0.5\sigma_r, \tag{16}$$

where σ_r is the smallest non-zero singular value of $K_{2,3}^{(l)}$. We then perform a SVD of $\widehat{K}_{2,3}^{(l)}$ and discard all singular values with value below the threshold $g^{\epsilon}(\nu^{(l)})$. Therefore, with probability at least $1-\delta$, the number of remaining singular values is equal to the true rank r. We are left with finding a suitable definition for the threshold function g^{ϵ}. From the condition on Eq. 16, we notice that we need g^{ϵ} to be smaller than a fixed value (RHS) and, at the same time, greater than a decreasing function of order $\mathcal{O}(\sqrt{\frac{1}{\nu^{(l)}}})$ (LHS). Then it is natural to define

$$g^{\epsilon}(\nu^{(k)}(l)) = \frac{g}{\nu^{(k)}(l)^{0.5-\epsilon}}$$

for a suitable $g > 0$ and with $0 < \epsilon < 0.5$. Therefore there is a number $N_0^{(l)}$ such that for all $\nu^{(l)} \geq N_0^{(l)}$ the condition on Eq. 16 is satisfied and Lemma 4 holds. Therefore we restate the sample complexity in Lemma 4 by adding the extra term to

$$\overline{N} \leftarrow \overline{N} + N_0(l).$$

Let $\overline{N}_{\max}$ denotes the maximum of this threshold for any action and policy.

C Proof of Lemma 1

Under policy π, Fig. 1-*right* shows the structure of $V_2^{(l)}$. Given action l, the matrix $V_2^{(l)}$ contains $X_{\pi}^{(l)}$ columns and each column corresponds to a column in emission matrix (up to permutation). We showed that the knowledge about a column of $V_2^{(l)}$ reveals part of the corresponding column in emission matrix, the entries with non-zero $\pi(y|l)$. The policy, in general, partitions the observation space to at most A partitions, $\mathcal{Y}_l \forall l \in \mathcal{A}$ and maps each partition to an action. It means that when we condition on an action, e.g., l, we restrict ourselves to the part of observation space $\mathcal{Y}_l$ and the input to the spectral learning algorithm is set $\mathcal{Y}_l$. Therefore, the algorithm is able to partition this set to $X_{\pi}^{(l)}$ partition. Because of the unknown permutation over columns of $V_2^{(l)}$ for different actions, we are not able to combine the resulting clustering give different actions. If we enumerate over actions, we end up with A partition $\mathcal{Y}_l$ and then we partition each set $\mathcal{Y}_l$ to at most X(upper bound on $X_{\pi}^{(l)}$), as a consequence, we might end up with at most XA disjoint clusters.

D Proof of Lemma 2

We first study the eigendecomposition of $M_2^{(l)}$ when its eigenvalues have multiplicity 1.

Lemma 5. *For any action $l \in [A]$, let the second moment in Eq. 3 have the eigendecomposition $M_2^{(l)} = U\Sigma U^{\mathsf{T}}$. If all eigenvalues of $M_2^{(l)}$ have multiplicity 1, there exists a mapping $\sigma^{(l)} : X \to X$ and multiplicative constants $\{C_i^{(l)}\}_{i\in[X]}$, such that for any $i \in \mathcal{X}_{\pi}^{(l)}$ and $j \in [Y]$, $[V_2^{(l)}]_{j,\sigma^{(l)}(i)} = C_i^{(l)}[U]_{j,i}$. As a result, for any hidden state $i \in \mathcal{X}_{\pi}^{(l)}$ we define the cluster $\widetilde{\mathcal{Y}}_i^{(l)}$ as*

$$\widetilde{\mathcal{Y}}_i^{(l)} = \{j \in [Y] : [U]_{j,i} > 0\} \tag{17}$$

and we have that if $j, j' \in \widetilde{\mathcal{Y}}_i^{(l)}$ then $j, j' \in \mathcal{Y}_{\sigma(i)}^{(l)}$ (i.e., observations that are clustered together in $\widetilde{\mathcal{Y}}_i^{(l)}$ are clustered in the original ROMDP).

In Eq. 3 we show that matrix $M_2^{(l)}$ is a symmetric matrix and has the following representation;

$$M_2^{(l)} := \sum_{i\in\mathcal{X}_{\pi}^{(l)}} \omega_{\pi}^{(l)}(i)[V_2^{(l)}]_{:,i} \otimes [V_2^{(l)}]_{:,i}.$$

As long as $V_2^{(l)}]_{:,i}$ for $i \in \mathcal{X}_\pi^{(l)}$ are orthogonal vectors, this matrix has rank of $X_\pi^{(l)}$ with the following eigendecomposition;

$$M_2^{(l)} = U\Sigma U^\mathsf{T}$$

where the matrix U, up to permutation, is the orthonormal version of $V_2^{(l)}$, and Σ is a diagonal matrix of rank $X_\pi^{(l)}$ with diagonal entries equal to $\omega_\pi^{(l)}$ multiplied by the normalization factors. As a result we can use the decomposition to directly recover the non zero elements of $V_2^{(l)}$ and the corresponding partial clustering.

Let's consider the i'th and j'th nonzero diagonal entries of matrix Σ, σ_i and σ_j, with eigenvectors of U_i, U_j, i.e., $M_2^{(l)}U_i = \sigma_i U_i$, $M_2^{(l)}U_j = \sigma_j U_j$. In the case of no eigengap, i.e., $\sigma_i = \sigma_j$, for any $0 \leq \lambda \leq 1$ we have $M_2^{(l)}(\lambda U_i + (1-\lambda)U_j) = \sigma_i(\lambda U_i + (1-\lambda)U_j) = \sigma_j(\lambda U_i + (1-\lambda)U_j)$. Therefore, any direction in the span of $span(U_i, U_j)$ is an eigenvector and the matrix decomposition is not unique, and we can not learn the true $V_2^{(l)}$. We relax this issue by deploying tensor decomposition of higher order moments.

The proof of Lemma 2 directly follows from the properties of tensor decomposition in [1] and the use of $V_2^{(l)}$ to generate a partial clustering.

E Proof of Theorem 1

The overall proof is mostly based on the original UCRL proof in [13]. In the following we refer to the exact steps in the original proof whenever we borrow results directly from it. The regret can be decomposed as follows;

$$\textit{Reg}_N = N\eta^* - \sum_{t=1}^{N} r_t = \sum_{k=1}^{K} \underbrace{\sum_{t=t_k}^{t_{k+1}-1} \big(\eta^* - \overline{r}(x_t, a_t)\big)}_{\Delta_k} + \sum_{t=1}^{N} \big(\overline{r}(x_t, a_t) - r_t\big),$$

where K is the total (random) number of episodes, x_t is the hidden state of the MDP at time t, same as r_t is the reward at time t, and $\overline{r}(x_t, a_t)$ is the true mean of reward. Using Heffting inequallity, as in Eq. 8 in [13], the last term can be bounded as $O(\sqrt{N\log(1/\delta)}$ with high probability. We then focus on the per-step regret Δ_k. At any epoch k, from Corollary 1 we know that any auxiliary state $s \in \widehat{\mathcal{S}}^{(k)}$ is a cluster of observations with same same hidden state. As a result, the reward $\overline{r}(x_t, a_t)$ is equivalent to $\overline{r}(y_t, a_t)$ (recall that all observations have the same reward as their hidden state). Therefore;

$$\Delta_k = \sum_{a, s \in \widehat{\mathcal{S}}^{(k)}} \nu^{(k)}(s, a)\big(\eta^* - r(s, a)\big).$$

The case when confidence intervals fail is bounded as in the original analysis. We now proceed with the same decomposition as done in [13] (Eqs.10,13,16) and obtain[7]

$$\Delta_k = \boldsymbol{\nu}^{(k)}\big(\widetilde{P}^{(k)} - P^{(k)}\big)w^{(k)} + \sum_{s \in \widehat{\mathcal{S}}^{(k)}, a} \nu^{(k)}(s, a)\big(\widetilde{r}^{(k)}(s, a) - \overline{r}(s, a)\big) + \boldsymbol{\nu}^{(k)}\big(I - P^{(k)}\big)w^{(k)},$$

where $\boldsymbol{\nu}^{(k)}$ is the vector of number of samples to auxiliary states in epoch k, $P^{(k)}$ (resp. $\widetilde{P}^{(k)}$) is the true (resp. optimistic) transition matrix over auxiliary states of policy $\pi^{(k)}$, $w^{(k)}$ is the centered version of the bias function returned by extended value iteration and $\widetilde{r}^{(k)}(s, a)$ is the optimistic reward. The first two terms account for the errors in estimating the dynamics and rewards of the (auxiliary) MDP and can be bounded as

$$\Delta_k \leq \Box D_{\widehat{\mathcal{S}}^{(k)}} \sqrt{\widehat{S}^{(k)} \log(1/\delta)} \sum_{s \in \widehat{\mathcal{S}}^{(k)}, a} \sqrt{N^{(k)}(s, a)} + \boldsymbol{\nu}^{(k)}\big(I - P^{(k)}\big)w^{(k)},$$

[7] Here we ignore the additive regret coming from approximate extended value iteration that accounts for an extra $O(\sqrt{N})$ regret at the end.

where $D_{\widehat{\mathcal{S}}^{(k)}}$ is the diameter of the auxiliary MDP at epoch k and $\square$ denotes universal numerical constants. The remaining term can be cumulatively bounded following similar steps as in Eq.18 in [13] with the only difference that the range of $w^{(k)}$ changes at each epoch. Thus we have

$$\sum_{k=1}^{K} \boldsymbol{\nu}^{(k)}\big(I - P^{(k)}\big)w^{(k)} \leq \square\sqrt{\sum_{k=1}^{K} D_{\widehat{\mathcal{S}}^{(k)}}\nu^{(k)}},$$

where $\nu^{(k)} = \sum_{s,a}\nu^{(k)}(s,a)$ is the length of epoch k. Grouping all the terms lead to the first regret statement

$$\mathit{Reg}_N \leq \square\bigg(\sum_{k=1}^{K} D_{\widehat{\mathcal{S}}^{(k)}}\sqrt{\widehat{S}^{(k)}\log(1/\delta)}\sum_{s\in\widehat{\mathcal{S}}^{(k)},a}\sqrt{N^{(k)}(s,a)} + \sqrt{\sum_{k=1}^{K} D_{\widehat{\mathcal{S}}^{(k)}}\nu^{(k)}}\bigg).$$

E.1 Clustering Rate

The first regret bound still contains random quantities in terms of the auxiliary MDPs generated over episodes. In this section we derive bounds on the number of steps needed to cluster observations. We notice that the analysis is extremely "pessimistic" and as we take worst-case values for all the quantities involved in the analysis.

Time to clustering. We proceed as follows. We first compute the minimum number of samples $\overline{N}(y)$ to guarantee that an observation y is correctly clustered. We then compute the length $\overline{\nu}(y)$ of an epoch so that $\overline{N}(y)$ samples are collected. Finally, we derive how many epochs $\overline{K}(y)$ are needed before an epoch of length $\overline{\nu}(y)$ is run.

We start by defining the probability that a certain action is explored. Let π be an arbitrary policy such that action a is taken in at least one observation y belonging to a hidden state $x = x_y$. Whenever an agent is in state x, there is a probability $\mathbb{P}(y|x)$ to observe y and thus trigger action a. We define the probability of "observing" an action a in state x under policy π as

$$\alpha_\pi(l) = \sum_{y\in x}\mathbb{P}(y|x)\mathbb{1}(\pi(y) = a). \tag{18}$$

Since we assumed that l is taken in at least one observation $\alpha_\pi(l)$ is always non-zero and it indeed lower-bounded by $O_{\min}$. We define $\alpha_p := \min_{k\in[K],x\in\mathcal{X},a\in\mathcal{A}}\alpha_{\pi^{(k)}}(a)$ as the worst proportion across all epochs, states, and actions.

Now we need need to know how fast the set $\mathcal{S}^{(k)}$ converges to set $\mathcal{X}$, in other work how fast is the clustering process. From Eq. 7 and the clustering process, we know that an observation y is clustered in x_y if the number of samples $\nu^{(k)}(l)$ obtained from the action l executed in y within a given epoch k is such that

$$f_O(y|x(y)) \geq 2C_O^{(k)}(l)\sqrt{\frac{\log(1/\delta)}{\nu^k(l)}}. \tag{19}$$

By reverting the bound and taking the worse case over actions and epochs, we obtain that a sufficient condition is to collect at least $\overline{N}(y)$ samples, with

$$\overline{N}(y) := \max_{l,k} 4C_O^{(k)}(l)\frac{\log(1/\delta)}{(f_O(y|x(y)))^2}.$$

We can now leverage on the ergodicity of the MDP and the probability of observation of an action α_p to find the minimum number of steps with an epoch to guarantee that with high probability the condition in Eq. 19 is satisfied. We define the worst-case mean returning time as $\tau_M = \max_\pi\max_x\mathbb{E}[\tau_\pi(x \to x)]$, where $\tau_\pi(x \to x)$ is the random time to go from x back to x through policy π. By Markov inequality, the probability that it takes more than $2\tau_M$ time step to from first visit of state x to its second visit is at most $1/2$. Given the definition of α_p, it is clear that if the action l is taken in state x then, this action will be taken at state x for α_p portion of the time. If we divide the episode of length ν into $\nu\alpha_p/2\tau_M$ intervals of length $2\tau_M/\alpha_p$, we have that within each interval we have a probability of $1/2$ to observe a sample from state x and take a particular action. Therefore, the lower bound on the average number of time that the agent takes any action l (that has a non zero probability to be executed in a state x) is $\nu\alpha_p/4\tau_M$ samples. Thus from Chernoff-Hoeffding, we obtain that the number of samples of any feasible action in the epoch with length ν is as follows;

$$\forall x \in \mathcal{X}, \forall l \in range\{\widetilde{\pi}(\cdot|x)\};\ \nu(l) \geq \frac{\nu\alpha_p}{4\tau_M} - \sqrt{\frac{\nu\alpha_p \log(XA/\delta)}{2\tau_M}}$$

with probability at least $1-\delta$. At this point, we can derive a lower bound on the length of the episode that guarantee the desired number of samples to reveal the identity of any observation is reached. For observation y, we solve

$$\frac{\nu\alpha_p}{4\tau_M} - \sqrt{\frac{\nu\alpha_p \log(XA/\delta)}{2\tau_M}} \geq \overline{N}(y)$$

and we obtain the condition

$$\sqrt{\nu} \geq \sqrt{\frac{2\tau_M}{\alpha_p}\log(XA/\delta)} + \sqrt{\frac{2\tau_M}{\alpha_p}\log(XA/\delta) + \frac{4\tau_M}{\alpha_p}\overline{N}(y)},$$

which can be simplified to

$$\nu \geq \overline{\nu}(y) := \frac{4\tau_M}{\alpha_p}\left(\overline{N}(y) + \log(XA/\delta)\right). \tag{20}$$

With the same argument in App. D in [4] the number of required epochs to reveal observation y is $\overline{K}(y) \leq AY\log_2(\overline{\nu}(y)) + 1$.

Time to clustering. Let $y_{\text{first}} = \arg\min_{y\in\mathcal{Y}} \overline{K}(y)$ be the first observation that could be clustered[8] then we define $K_{\text{first}} = \overline{K}(y_{\text{last}})$ as the number of episodes and $N_{\text{first}} = 4AY\overline{\nu}(y_{\text{first}})$ the total number of steps needed before clustering y_1 correctly. Similarly, let $y_{\text{last}} = \arg\max_{y\in\mathcal{Y}}$ and $K_{\text{last}} = \overline{K}(y_{\text{last}})$ as the number of episodes and $N_{\text{last}} = 4AY\overline{\nu}(y_{\text{last}})$ the total number of steps needed before clustering y_{last} correctly. Since all the other observations will be clustered before y_{last} we can say that by epoch K_{last} all observations will be clustered. As discussed in Lem. 1 this does not necessarily correspond to the hidden state $\mathcal{X}$ but it could be an auxiliary space $\mathcal{S}$ with at most AX states.

Validity of the bound in Lemma 3. We also notice that Lemma 3 requires a minimum number of samples N_0 before the concentration inequality on the estimate of V_2 holds. Applying a similar reasoning as for the time for clustering, we can derive a bound on the number of episodes K_{sm} and number of samples N_{sm} needed before the spectral method actually works (from a theoretical point of view). As a result, a more accurate definition of K_{first} and K_{last} (resp. for N) should take the maximum between the values derived above and K_{sm}.

E.2 Minimal Clustering

The spectral learning algorithm has been shown to efficiently cluster the observation set to an auxiliary state space of size $X \leq S \leq XA$. As long as different clusters are merged across epochs, we expect $\mathcal{S}^{(k)}$ to tends to $\mathcal{X}$, yet there is a chance that it converges to a number of auxiliary states $S \neq X$. To make sure that the algorithm eventually converges to the hidden space $\mathcal{X}$, we include a further clustering technique. We adapt the idea of Gentile et al. [9], Cesa-Bianchi et al. [7] and the state aggregation analysis of Ortner [21] and perform an additional step of *Reward and Transition Clustering*. In order to simplify the notation, in the following we remove the dependency on k, even if all the quantities should be intended as specifically computed at the beginning of epoch k.

We first recall that given any hidden state x and any action a we have $r(y,a) = r(x,a)$ (*reward similarity*) and $p(\cdot|y,a) = p(\cdot|x,a)$ for all observations $y \in \mathcal{Y}_x$ (*transition similarity*). The same similarity measures work for auxiliary states via replacing observations with auxiliary states in the above definitions, i.e., given any hidden state x and any action a we have $r(s,a) = r(s,a)$ and $p(\cdot|s,a) = p(\cdot|x,a)$ for all auxiliary states $s \in \mathcal{S}$ that belong to hidden state x.[9] We also recall that

[8] This should be intended as the first observation that is clustered in the worst case. In practice, depending on the policy and the randomness in the process, other observation may actually be clustered well before y_1.

[9] Notice that this holds since $\mathcal{S}$ is a "valid" clustering in high probability.

high-probability confidence intervals can be computed for any $s \in \widehat{\mathcal{S}}$ any $a \in \mathcal{A}$ as

$$\begin{aligned}\|p(\cdot|s,a) - \widehat{p}(\cdot|s,a)\|_1 \leq d(s,a) &:= \sqrt{\frac{14\widehat{S}\log(2AN/\delta)}{2\max\{1,N(s,a)\}}} \\ |\bar{r}(s,a) - \widehat{r}(s,a)| \leq d'(s,a) &:= \sqrt{\frac{7\log(2\widehat{S}AN/\delta)}{2\max\{1,N(s,a)\}}}.\end{aligned} \tag{21}$$

At any epoch, we proceed by merging together all the auxiliary states in $\widehat{S}$ whose reward and transition confidence intervals overlap (i.e., s and s' are merged if the confidence interval $[\widehat{r}(s,a) \pm d_r(s,a)]$ overlaps with $[\widehat{r}(s',a) \pm d_r(s',a)]$ and $[\widehat{p}(\cdot|s,a) \pm d_p(s,a)]$ overlaps with $[\widehat{p}(\cdot|s',a) \pm d_p(s',a)]$) and construct a new set $\widetilde{\mathcal{S}}$. In practice, the set $\widetilde{S}$ is constructed by building a fully connected graph on $s \in \widehat{S}$ where each state s as a node. The algorithm deletes the edges between the nodes when $|\widehat{r}(s,a) - \widehat{r}(s',a)| > d_r(s,a) + d_r(s',a)$ or $|\widehat{p}(\cdot|s,a) - \widehat{p}(\cdot|s',a)|_1 > d_p(s,a) + d_p(s',a)$. The algorithm temporarily aggregates the connected components of the graph and consider each disjoint component as a cluster. If the number of disjoint components is equal to X then it returns $\widehat{\mathcal{S}}$ as the final hidden state $\mathcal{X}$, otherwise the original auxiliary state space $\widehat{\mathcal{S}}$ is preserved and the next epoch is started. Notice that if s and s' belong to the same hidden state x then w.h.p. their confidence reward and transition intervals in Eq. 21 overlap at any epoch. Thus in general $\widetilde{S} \leq X$.

Let's define the reward gaps as follows (similar for the transitions) $\forall s, s' \in \widehat{\mathcal{S}}, \forall a \in \mathcal{A}$ and the corresponding x, x'

$$\begin{aligned}\gamma_r^a(s,s') = \gamma_r^a(x,x') &:= \big|\bar{r}(x,a) - \bar{r}(x',a)\big| = \big|\bar{r}(s,a) - \bar{r}(s',a)\big|, \\ \gamma_p^a(s,s') = \gamma_p^a(x,x') &:= \|p(\cdot|x,a) - p(\cdot|x',a)\|_1 = \|p(\cdot|s,a) - p(\cdot|s',a)\|_1.\end{aligned}$$

where $p(\cdot|x,a), p(\cdot|s,a) \in \widetilde{\Delta}_{\widehat{S}-1}$, where $\widetilde{\Delta}_{\widehat{S}-1}$ is $(\widehat{S}-1)$ dimensional simplex. To delete an edge between two states s, s' belonging to two different hidden states, one of the followings needs to be satisfied for at least for one action

$$|\widehat{r}(s,a) - \widehat{r}(s',a)| > d_r(s,a) + d_r(s',a) \Rightarrow \gamma_r^a(s,s') > \sqrt{\frac{7\log(2\widehat{S}AN/\delta)}{2\max\{1,N(s,a)\}}} + \sqrt{\frac{7\log(2\widehat{S}AN/\delta)}{2\max\{1,N(s',a)\}}} \tag{22}$$

$$|\widehat{p}(\cdot|s,a) - \widehat{p}(\cdot|s',a)|_1 > d_p(s,a) + d_p(s',a) \Rightarrow \gamma_p^a(s,s') > \sqrt{\frac{14\widehat{S}\log(2AN/\delta)}{2\max\{1,N(s,a)\}}} + \sqrt{\frac{14\widehat{S}\log(2AN/\delta)}{2\max\{1,N(s',a)\}}} \tag{23}$$

For simplicity, we proceed the analysis with respect to reward, the same analysis holds for transition probabilities. The Eq. 23 can be rewritten as follows;

$$\left(\frac{1}{\sqrt{\max\{1,N(s,a)\}}} + \frac{1}{\sqrt{\max\{1,N(s',a)\}}}\right)^{-1} \geq \sqrt{\frac{28\log(2\widehat{S}A^2N/\delta)}{2\gamma_r^a(s,s')^2}}$$

which hold when

$$\min\{N(s,a),N(s',a)\} > \frac{56\log(2\widehat{S}AN/\delta)}{\gamma_r^a(s,s')^2}. \tag{24}$$

This implies that after enough visits to the auxiliary states s and s', the two states would be split if they belong to different hidden states. We notice that as $\mathcal{S}$ becomes smaller, more and more samples from raw observations are clustered into the auxiliary states, thus making $N(s,a)$ larger and larger. Furthermore, we can expect that the transition gaps may become bigger and bigger as observations are clustered together.

The way that spectral method clusters the observation is effected by separability of observations' probability. But the clustering due to reward analysis (or transition or both) is influenced by the separability in reward function (or transition function or both) and depends on gaps. These two methods look at the clustering problem from different point of view, as a consequence, their combination speeds up the clustering task.

For simplicity we just again look at the reward function, same analysis applies to transition function as well.

Regret due to slowness of Reward Clustering (Transition Clustering) Let $N_a^r(s,s')$ denote the required number of sample for each of s and s' to disjoint them.

$$N_a^r(s,s') := \frac{56\log(2\widehat{S}AN/\delta)}{\gamma_a(s,s')^2}$$

While the underlying Markov chain is ergodic, with high probability we can say at time step $N(s,s')$, at least for one action $\min\{N(s,a),N(s',a)\} \leq \overline{N}_a(s,s')$ where

$$N_r(s,s') = \min_a\{\frac{4\tau_M}{\alpha_p}\left(N_a^r(s,s') + \log(YA/\delta)\right)\}$$

In the worse case analysis we might need to have $N_r = \max_{s,s'} N_r(s,s')$ samples, which corresponds to at most $AY\log(N_r)$ episode. At this time, the reward clustering procedure can output the exact mapping. This bound can be enhance even further by considering the reward function together with the transition process. With the same procedure we can define $N_p = \max_{s,s'} N_p(s,s')$ where

$$N_p(s,s') = \min_a\{\frac{4\tau_M}{\alpha_p}\left(N_a^p(s,s') + \log(YA/\delta)\right)\} \tag{25}$$

with

$$N_a^p(s,s') := \frac{112\widehat{S}\log(2\widehat{S}AN/\delta)}{\gamma_a(s,s')^2}$$

Again, in the worse case analysis we might need to have $N_p = \max_{s,s'} N_p(s,s')$ samples, which corresponds to at most $AY\log(N_r)$ episode. Therefore the number of required episode for the agent to declare the true mapping w.h.p., is $AY\log(\min\{N_r,N_p\})$.

F Extended Discussion

Conjecture 1. *We can extend this results to deterministic MDP. In this case we can first uniformly explore the latent space and collect sufficient number of sample to find the exact clustering and reduce the large MDP to the latent MDP and then apply* UCRL *on the latent MDP. Which can suffer a constant regret of pure exploration at the beginning and regret of $\widetilde{\mathcal{O}}(D_{\mathcal{X}}X\sqrt{AN})$ due to* UCRL *in the second phase. One of the main open questions is whether the spectral clustering method could still provide "useful" clusterings when the state space is not fully visited (i.e., in case of non-ergodic MDP), so that observations are properly clustered where it is actually needed to learn the optimal policy. We can provide a partial answer in the case of deterministic ROMDPs. In fact, despite not being ergodic, in this case we can first uniformly explore the latent space and collect sufficient number of sample to find the exact clustering and reduce the large MDP to the latent MDP and then apply* UCRL *on the latent MDP. These two-phase algorithm would suffer a constant regret of pure exploration at the beginning and regret of $\widetilde{\mathcal{O}}(D_{\mathcal{X}}X\sqrt{AN})$ due to* UCRL *in the second phase.*

In RL problems, the principle of Optimism-in-Face-of-Uncertainty contributes in designing a policy that locally improves the model uncertainty and average reward which has been shown to be an optimal strategy. It is an open question to analyze and modify this principle for the models with clustering where global improvement of the information in model uncertainty is required. While the SL-UC for deterministic models reaches order optimal regret, it is not still clear how to modify the exploration to enhance the constant regret of the pure exploration phase.

Compared to POMDP. We can compare this result with the regret bound of Azizzadenesheli et al. [4] for POMDPs, which are a more general class than ROMDPs. Recalling that POMDPs are characterized by a diameter $D_{pomdp} := \max_{x,x',a,a'}\min_{\pi\in\mathcal{P}}\mathbb{E}[\tau\left((x,a)\rightarrow(x',a')\right)]$, the regret derived by Azizzadenesheli et al. [4] scales as $\widetilde{\mathcal{O}}(D_{pomdp}X^{3/2}\sqrt{AYN})$. The regret suffers from additional term $\sqrt{Y}$ because the RL algorithm in POMDP put much effort on accurate estimation of entries of O matrix and does not exploit its specific structure. Moreover, there is an additional factor

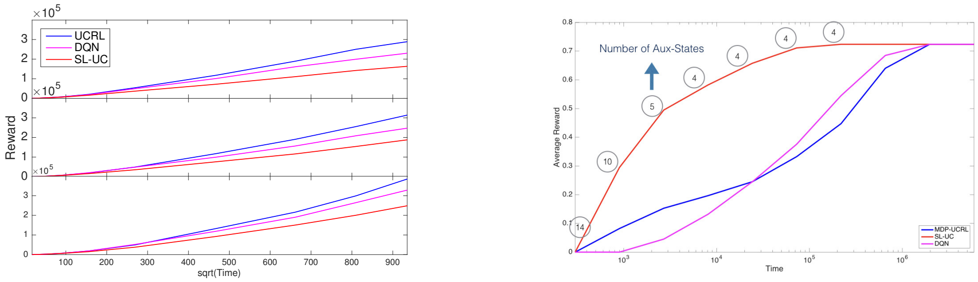

Figure 3: *(left)*The regret comparison, $A = 4$, from top to bottom, $Y = 10, 20, 30$. The scale is $\sqrt{T}$. *(right)* Learning rate of SL-UC compared to UCRL and DQN. After first few rounds, it learns the the true mapping matrix. The numbers in the bulbs are the cardinality of Aux-MDP.

X in regret bound due to learning of transition tensor through spectral methods. Their experimental results are provided here [5].

G Additional Experiments

While the results reported in the main text are obtained on actual ROMDPs, here we test SL-UC on random MDPs with no explicit hidden space. The objective is to verify whether SL-UC can be used to identify (approximate) clusters. Since SL-UC in high probability only clusters observations that *actually* belong to the same hidden state, in this case SL-UC would reduce to run simple UCRL, as there is no two observations that can be *exactly* clustered. In order to encourage clustering, we half the (exact) confidence intervals in the attempt of trading off a small bias with a significant reduction in the variance. We compare the regret on three random MDPs with increasing number of states. As it is shown in Fig. 3, SL-UC is effective even in this scenario compared to UCRL and DQN. In fact, we see from Fig. 3-right that SL-UC is able to find clusters without compromising the overall regret. While the number of states now directly affects the performance of SL-UC, we see that it is more robust than the other algorithms and its regret is not severely affected by an increasing number of observations.